**SANDIA REPORT**

SAND2005-5262
Unlimited Release
Printed August 2005

# An Analysis of HyperTransport and Seastar Data Rates on Red Storm

Douglas W. Doerfler

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

**NOTICE:** This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2005-####
Unlimited Release
Printed August 2005

# An Analysis of HyperTransport and Seastar Data Rates on Red Storm

Douglas W. Doerfler
Scalable Systems Integration
Sandia National Laboratories
P.O. Box 5800, MS-0817
Albuquerque, New Mexico 87185

## Abstract

The Red Storm high-speed network system chip, Seastar, uses the HyperTransport (HT) interface to communicate with an on-board Opteron processor. An analysis is presented for a) the theoretical maximum sustained data rate for an 800 MHz, 16-bit HT implementation, b) the theoretical maximum sustained data rate for the Seastar link interface, and c) the measured maximum sustained bandwidths from readily accessible HT platforms, which include Red storm, Cray's XD1 and a Hewlett-Packard DL145 dual-Opteron server.

## Contents

## Tables

# 1.0 Introduction

The Red Storm high-speed network system chip, Seastar, uses the HyperTransport (HT) interface to communicate with an on-board Opteron processor. The interface is an 800 MHz, 16-bit wide implementation of HT[1]. An analysis is presented for a) the theoretical maximum sustained data rate for an 800 MHz, 16-bit HT implementation, b) the theoretical maximum sustained data rate for the Seastar link interface, and c) the measured maximum sustained bandwidths from readily accessible HT platforms, which include Red storm, Cray's XD1 and a Hewlett-Packard (HP) DL145 dual-Opteron server.

# 2.0 Peak and Theoretical Sustained Transfer Rates

## 2.1 Seastar

The Seastar link interface utilizes 12 LVDS signals, per direction. Each signal transfers data at a rate of 1.6 Gb/s. The links utilize double data rate signaling and hence each clock cycle delivers 2 bits. The links employ 8b10b encoding. After encoding, the data rate for the Seastar link is 3.84 GB/s per direction. The Seastar protocol breaks messages down into packets, packets are broken down into flits and flits are broken down into micropackets. A micropacket is sent over the link in a unit called a phit. The Seastar messaging protocol is:

- A message can be of arbitrary length.
- A packet consists of a header of size 1 flit, followed by up to 8 data flits.
- A flit is 68 bits in length, 4 bits control plus 64 bits of data.
- A micropacket is of size 2 flits, plus 32 bits of control for a total size of 168 bits.
- Micropackets are sent over the link 24 bits at a time, which is called a phit. The time to transfer a phit is one clock cycle. Hence, each micropacket takes 7 clock cycles to transmit.

Doing the math, the maximum sustained application data rate for a Seastar link is 2.60 GB/s per direction. This is 67.7% of the Seastar's raw data rate.

## 2.2 HyperTransport

The signaling rate of an 800 MHz HT implementation is 1,600 MT/s. With the 16-bit wide HT channel used on Red Storm, the HT link's raw data rate is 3.2 GB/s per direction. The actual data rate an application would achieve is limited by the HT protocol and it also depends on many design details. Important design factors include:

- The designs HT request buffer size and depth is critical on both the transmitting and receiving ends of a link.
- The ability of the HT device to move the data out of the request buffers to their destination devices.
- An HT link must support multiple virtual channels. In addition, it multiplexes address, data, and control for each virtual channel. Hence, flow control and traffic from other virtual channels can take bandwidth away from any given application data transfer.

The HT packet protocol for a read or write has a maximum data payload of 64 bytes per packet and this is preceded by an 8 byte request packet. For a uni-directional transfer and assuming an ideal condition in which each of the issues noted above are not a factor and the receiver is able to

[1] 1 MB/s = $10^6$ bytes/sec; 1 GB/s = $10^9$ bytes/sec; 1 MT/s = $10^6$ transfers/sec

accept a continuous stream of HT requests, the maximum sustained application data rate for an 800 Mhz, 16-bit HT link is 2.844 GB/s. This is 88.9% of the links raw data rate.

For a bidirectional transfer the flow control traffic for the opposite direction of the link will be competing with data transfers. Again assuming ideal conditions, this means a 4 byte flow control packet would be injected into the application's data stream for every 3 HT read or write requests. Hence, for every 3 HT requests of maximum application data size 64 bytes + 8 control bytes, the link must also carry a 4 byte flow control packet. Hence the efficiency for bidirectional transfers is (3*64) / (3*(64+8)+4) = 87.3% or 5.59 GB/s.

Based on the above analysis[2], in a transfer from user memory on one node to user memory on another node, the bottleneck from a theoretical perspective is the Seastar link.

## 3.0 Measured Sustained Transfer Rates

Large data, continuous streaming, data rate measurements were made for three platforms: Red Storm, Cray's XD1 and HP's dual-Opteron DL145 server. For Red Storm and the XD1, the uni-directional measurements were made using the Pallas PingPong MPI benchmark and the bi-directional measurements were made using the Pallas Sendrecv MPI benchmark. For the DL145 server, a modified version of the Streams Benchmark was used to measure the rate for one processor to read a large array of data from memory associated with the second onboard Opteron processor. For bi-directional measurements, each processor read remote memory simultaneously.

## 3.1 Red Storm and Seastar

The large message uni-directional and bi-directional measured MPI bandwidth for Red Storm is 1,160 MB/s and 2,160 MB/s respectively. It can be assumed that this is representative of the low-level messaging rates because for very large messages the Portals and MPI overheads are negligible. This measured uni-directional data rate is 40.8% of HT's and 44.6% of Seastar's theoretical maximum sustained application data rates. The respective measured bi-directional data rates are 38.7% of HT's and 41.5% of Seastar's theoretical maximums.

## 3.2 Hewlett-Packard DL145 with 800Mhz HyperTransport

Nodes used on Sandia's Red Squall Cluster, HP's dual-Opteron DL145, use HT to implement a non-uniform memory architecture (NUMA), 2-way server. A modified version of the Streams Memory Benchmark was used on one of Red Squall's compute nodes to measure the bandwidth observed by the application when a large array was read from memory to a local variable. When reading data from the remote processors memory, this provides a continuous, uni-directional data stream traveling over the HT link, similar to that observed in a uni-directional MPI messaging benchmark. In the application benchmark, the array is first filled from the processors local memory, and then from the memory attached to the remote processor. For bi-directional transfers, both processors simultaneously performed large array reads from the remote processors memory into a local memory location.

The average streaming bandwidth seen by the application when accessing it's local memory (i.e. HT is not involved) was 3.55 GB/s. When streaming memory from the remote processors memory (in this case HT is used as the transport) the average bandwidth achieved was 2.35 GB/s. The

[2] In addition, all of the systems tested in this study used PC2700 (333 MHz DDR) DRAM. The Opteron employs two channels (2 x 8 bytes wide), hence the peak DRAM data rate is 5.328 GB/s.

remote memory access bandwidth of 2.35 GB/s is 82.6% of HT's theoretical maximum sustained application data rate. The average bi-directional bandwidth achieved was 3.99 GB/s, or 71.4% of the theoretical maximum.

## 3.3 Cray XD1 with 400Mhz HyperTransport

The measured, large message, MPI bandwidth of Cray's XD1 platform is 1,270 MB/s uni-directional and 2,150 MB/s bi-directional. However, the XD1 uses a 400 MHz, 16-bit implementation of HT. Hence its theoretical maximum sustained HT bandwidth is half of Red Storm's HT implementation, 1.42 GB/s uni-directional and 2.79 GB/s bi-directional. The XD1's uses two 4x InfiniBand links between nodes with a theoretical peak bandwidth of 2 GB/sec per direction. Measurements performed on InfiniBand based clusters at Sandia have shown that the MPI efficiency for large messages is approximately 95% for uni-directional transfers and 85% for bi-directional transfers. So it's relatively safe to assume that the limiting factor in moving data between nodes on the XD1 is HT. The interesting observation is that MPI messaging on the XD1 achieves 89.3% uni-directional and 77% bi-directional of HT's theoretical maximum sustained transfer rates. Hence, it's reasonable to expect other HT implementations, including Red Storm's, to achieve similar efficiencies.

## 4.0 Summary and Conclusions

Based on the XD1 and DL145 analysis, Red Storm's expected sustained HT application data rate should be somewhere between 82% and 89% uni-directional and 71% to 77% bi-directional of the theoretical maximum sustained rate. Assuming 85% uni-directional and 75% bi-directional, the respective expected rates are 2.4 GB/s and 4.2 GB/s.

The Seastar link level layer should achieve 95% uni-directional and 85% bi-directional of its theoretical maximum sustained application data rate, based on observations made on other HPC platforms at Sandia. Hence, the expected sustained transfer rate of the Seastar link is 2.60 GB/s x 0.95 = 2.47 GB/s unidirectional and 5.2 GB/s x 0.85 = 4.42 GB/s.

The expected data rates for Red Storm's HT and Seastar interfaces are approximately equal. So it is difficult to predict which link will be the bottleneck in messaging applications. However, this assumes that issues can be resolved to bring the measured data rates close to those expected.

Table 1: Peak, Theoretical Sustained, and Measured Transfer Rates

| | Peak BW (GB/s) | | | |
|---|---|---|---|---|
| | uni-directional | bi-directional | | |
| Seastar | 3.84 | 7.68 | | |
| 800Mhz HT | 3.2 | 6.4 | | |
| 400Mhz HT | 1.6 | 3.2 | | |
| | Theoretical Sustained BW (GB/s) | | % of Peak BW | |
| | uni-directional | bi-directional | uni-directional | bi-directional |
| Seastar | 2.6 | 5.2 | 67.7% | 67.7% |
| 800Mhz HT | 2.84 | 5.59 | 88.9% | 87.3% |
| 400Mhz HT | 1.42 | 2.79 | 88.9% | 87.3% |
| | Measured Sustained BW (GB/s) | | % of Theoretical Sustained HT BW | |
| | uni-directional | bi-directional | uni-directional | bi-directional |
| Seastar: Red Storm [1] | 1.16 | 2.16 | 40.8% | 38.7% |
| 800Mhz HT: DL145 | 2.35 | 3.99 | 82.6% | 71.4% |
| 400 Mhz HT: XD1 [1] | 1.27 | 2.15 | 89.3% | 77.0% |
| | Expected Sustained BW (GB/s) [3] | | Expected % of Theoretical Sustained BW | |
| | uni-directional | bi-directional | uni-directional | bi-directional |
| Seastar: Red Storm [2] | 2.4 | 4.2 | 85% | 75% |

1) Assumes HT is the limiting factor

2) Expected percentages are based on XD1 results

## 5.0 References

1. Jay Troden and Don Anderson, *HyperTransport System Architecture*, MindShare, Inc., Addison-Wesley, 2003.
2. "Red Storm Seastar System Chip Specification", Cray Inc., Version 1.6, January 9th, 2004.
3. Douglas Doerfler, "A Quick Look at InfiniBand: November 2004", Sandia National Laboratories, SAND2005-1555P, March, 2005.

## Distribution:

| | | | |
|---|---|---|---|
| 1 | MS-0316 | Sudip Dosanjh | 01420 |
| 1 | MS-0321 | Bill Camp | 01400 |
| 1 | MS-0376 | Ted Blacker | 01426 |
| 1 | MS-0816 | Jim Ang | 01422 |
| 1 | MS-0822 | David White | 01427 |
| 1 | MS-1109 | Jim Tomkins | 01420 |
| 1 | MS-1110 | Neil Pundit | 01423 |
| 1 | MS-9018 | Central Technical Files | 8945-1 |
| 2 | MS-0899 | Technical Library | 9616 |